DOCUMENT RESUME

ED 463 523 CS 510 848

AUTHOR Rhodes-Kline, Anne K.
TITLE Estimating Stanine Scores from a Non-Random Sample: A Methodology Discussion.
PUB DATE 1997-05-01
NOTE 17p.; Paper presented at the Annual Meeting of the New England Educational Research Organization (Portsmouth, NH, April 30-May 2, 1997). Formerly titled, "Literacy Acquisition into the Twenty-First Century: Stanines Now and a Method for the Future."
AVAILABLE FROM The University of Maine, College of Education & Human Development, 5766 Shibles Hall, Orono, ME 04469-5766; Tel: 207-581-2438; Fax: 207-581-2423; Web site: http://www.ume.maine.edu/~cel.
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Early Intervention; Primary Education; *Reading Achievement; Reading Improvement; *Research Methodology; Sample Size; Sampling; Statistical Analysis; *Statistics; *Student Evaluation
IDENTIFIERS Maine; *Reading Recovery Projects

ABSTRACT

A methodology for estimating descriptive statistics, specifically the mean and the variance, from a sample that is not normally drawn is described. The method involves breaking the sample down into subgroups and weighting the descriptive statistics associated with each subgroup by the proportion of the population that the subgroup represents. This methodology is demonstrated using a data set of almost 4,000 children in Maine associated with the Reading Recovery program. After the calculation of means and variances, stanine scores for the spring administration of six literacy assessments are computed. (Author/RS)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 463 523

Estimating Stanine Scores from a Non-Random Sample

A Methodology Discussion

(Formerly titled,

"Literacy Acquisition into the Twenty-first Century:

Stanines Now and a Method for the Future")

NEERO CONFERENCE PAPER

May 1, 1997

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

T F Moore

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Anne K. Rhodes-Kline, Ph.D.

5766 Shibles Hall

University of Maine

Orono, ME 04469

CS 510 848

BEST COPY AVAILABLE

Abstract

A methodology for estimating descriptive statistics, specifically the mean and the variance, from a sample that is not normally drawn is described. The method involves breaking the sample down into subgroups and weighting the descriptive statistics associated with each subgroup by the proportion of the population that the subgroup represents. This methodology is demonstrated using a data set of almost four thousand children in Maine associated with the Reading Recovery program. After the calculation of means and variances, stanine scores for the spring administration of six literacy assessments are computed.

Stanine Scores from a Reading Recovery Data Set
Maine Norms for the Observation Survey

Education is heading into the twenty-first century in a climate of accountability. President Clinton has called for every third grader to be able to read (Clinton, 1997). Educators are working now in every state to establish learning standards for children at each grade level.

Statewide norms can be useful to K-2 educators in making assessment decisions about children, and they can also be useful for making school and district-level programmatic decisions. This paper describes a method for estimating descriptive statistics, specifically the mean and the variance, when a simple random sample from the population is not available. Specifically, it details the application of this method to the data set associated with the Reading Recovery program in Maine. The method should generalize to other samples in which the sizes of the population subgroups are known. The descriptive statistics estimated are then used to compute stanine scores.

Reading Recovery Program

Reading Recovery is a one-on-one early-intervention program for first graders at risk of literacy failure. To determine eligibility for the program, the Kindergarten teacher is asked to submit, at the end of the year, a list of the children he or she thinks are behind the others in literacy learning (on skills such as letter knowledge and print awareness). These children are tested using the Observation Survey (described below), those with the lowest scores are taken into the program first. The remaining children recommended for the program are placed on a waiting list. As the neediest children acquire the skills that will enable them to progress at the level of their classmates, they are released, and children from the waiting list can begin in their places.

In Maine, data on literacy skills are collected in both the fall and the spring of the first grade year on all children who are recommended for Reading Recovery and on a random sample of the

children in each school who are not recommended or later identified. The latter group, called the "random sample," provides a target or goal point for the skills of children in Reading Recovery.

Observation Survey

The Observation Survey is a battery of six reading and writing assessments for children who are just beginning to acquire literacy skills. It was developed in New Zealand by Clay (1993) to assess the literacy skills of young children and to identify children who are at risk of not learning to read and write. The Observation Survey has been adapted for the U.S. context and has achieved wide use in the U.S. (Reading Recovery Council of North America, 1993) through its association with the Reading Recovery program.

For the Letter Identification task, children are shown a page of randomly arranged letters, including all lower and upper case letters (printed) as well as the letters g and a in a typewritten font. Children are given credit for a correct response if they name the letter, give an appropriate letter sound, or name a word that starts with the letter. The maximum score on this assessment is 54.

The Concepts About Print assessment in the Observation Survey measures a group of behaviors that reveal what children understand about the conventions of written language. Areas measured include knowing how to open a book; knowing when pictures and print are right-side up or upside down; knowing where to read when a page has both text and picture; and knowing when letters or words are out of order. Scores on this assessment range from 0 to 24.

The Writing Dictation task is a measure of how well a child can hear and record sounds. The observer selects one of five possible sentences to dictate. He/she encourages the child to say each word slowly and write the sounds heard. Performance is measured by the number of phonemes correctly represented, even though the word may not be correctly spelled (i.e., toda for tod.)

When a child reverses the order of phonemes when representing the sounds (i.e., ma for am), one point is subtracted from the total score for each reversal. Additional letters do not affect the scoring (i.e., todae for today still scores four phonemes). The maximum possible score on the Writing Dictation task is 37.

The Word Test is a measure of single word recognition involving a list of twenty high frequency words. The child is asked to read the list, and each word correctly read (and pronounced) is scored as one point.

The Writing Vocabulary assessment measures the number of (different) words a child can write correctly in ten minutes. Children are prompted to write words they know. Performance is measured by the number of words written correctly.

The Text Reading measure of the Observation Survey assesses the highest gradient of text difficulty that the child can read with 90% accuracy or better. Accuracy is measured by the percentage of words correctly read and pronounced. All Reading Recovery schools use the *Scott Foresman Testing Packet* (1979), a standard set of benchmark-leveled texts. To administer the text assessment, the observer gives the child a brief introduction to a book and then asks the child to read the book independently. When a child scores below 90% accuracy on two gradients of text in a row, the observer discontinues the assessment. Scores range from 0, indicating an inability to read "no, no, no") to 30, indicating an ability to read more complex story structure, with fewer pictorial clues and more lines to a page. Text levels include all the integers from 1 to 10 and the even integers from 12 to 30. According to the publishers, level 30 is equivalent to a sixth grade reading level.

The Observation Survey spans the range of early literacy skills from preschool (age 3) to the end of first grade (age 7). For most children in preschool, the Letter Identification task and the

Concepts About Print task are useful. The other four measures tap more advanced early literacy skills, beyond the abilities of most preschoolers. At the end of Kindergarten and at entry to first grade, all six measures are useful, as most children this age have some but not all of the skills these tests measure. By the end of first grade, however, the Concepts About Print and Letter Identification tasks are no longer very useful, since most children in this age group score very high on both measures. It is expected that children will eventually master the skills that are measured by the six tests, so the Observation Survey is most useful as an assessment for only the years during which children are acquiring early literacy skills.

The Need for Stanine Scores

Most practitioners are interested in assessment for two primary purposes. One is measurement of an individual's growth and learning over time. For this purpose, assessments answer the question, "how much progress has the student made?" The student is compared to him- or herself at an earlier time. The other primary purpose for assessment is measuring a student's learning in relation to his or her peers. For this purpose, assessments can answer questions such as, "how well has this student done in relation to other children his or her age?" or, "how well has this classroom or school done in relation to the district or the state?"

In the fall, there are two questions about the new first grade class that practitioners answer with the Observation Survey. The first is, which entering first grade children have the lowest literacy skills? The second is, how does our Kindergarten program compare to others in the state, or, in other words, is it doing its job of preparing children for first grade?

In the spring, there are two questions about the exiting first grade class that practitioners answer with the Observation Survey. First, are the skills of the children who were served with the

Reading Recovery program within the average range of skills the other children (who were not served by the program)? Second, how adequate is our school's first grade program in relation to the other first grades in the state?

Educators in Maine like to use stanine scores, largely because they are easy to understand and familiar to almost all teachers and administrators across the state. Although the Observation Survey is used in about 75% of the elementary schools in Maine, until recently, the only norms that were available for the Observation Survey were a set of stanines normed on 160 children from urban Columbus, Ohio, in 1990. These stanines were an inadequate reference point for the scores of Maine children. Practitioners, especially those associated with the Reading Recovery program, were very eager for a set of stanines based on Maine children. This need was the impetus for the project whose methodology is the subject of this paper.

Data

One of the distinguishing characteristics of the Reading Recovery program is its emphasis on quantitative assessment and evaluation at both the individual child level and the regional and state levels. Observation Survey data are collected on children who enter the program, as well as on a random sample of children judged not to be at risk (and not eligible for the program). In Maine, Observation Survey data are also collected on children who are placed on the waiting list for the program but never served.

Effectively, these three groups of children (Reading Recovery children, waiting list children, and random sample children) can be thought of as three strata from the population of first graders who attend schools where Reading Recovery is implemented. (Reading Recovery is implemented in about 75% of the schools in Maine). The groups are mutually exclusive and exhaustive, since every

first grader falls into one and only one category. However, it should be pointed out that children in schools that do not have Reading Recovery might not fall clearly into one group or another.

Data were collected from every Reading Recovery child (1709), every waiting list child (755), and a sample (1450) of the others. So, population information is known about two of the three strata. For the third stratum, smaller schools (fewer than 20 first graders) were asked to collect data on the entire first grade class. Therefore, small schools' data represent the population for all three strata. For practical reasons, larger schools were not asked to collect data on every child, although some of these schools chose to do so as a more accurate determination of which children should receive Reading Recovery. Larger schools that chose not to test all first graders were asked to randomly select for testing eight students from the population of children judged not eligible for the program. The resulting state data set is unique in its composition, and no methodology for the estimation of norms from such a data set could be found.

The data include fall and spring scores on the six literacy assessments that comprise the Observation Survey. The data also include identifying information about the child's school and district. Additional data were collected at the district level regarding the size of each first grade class.

Methods

First, the data were aggregated by district and by category of child to form subgroups (each district theoretically could have three groups of children, although not every district had a waiting list). The proportion of the population that each subgroup represented was then determined. Since the populations of Reading Recovery and waiting list children were known (they were equal to their

sample sizes), the population of not-judged-to-be-at-risk children could be deduced by subtracting the sizes of the other two groups from the district's first grade population. Within each district,

$$P_{nar} = P_{tot} - P_{RR} - P_{wl} \quad , \qquad \text{(Equation 1)}$$

where P_{nar} is the population of children not judged to be at risk for literacy failure, P_{tot} is the first grade population in the district among schools with Reading Recovery, P_{RR} is the number of at-risk first graders served in the Reading Recovery program, and P_{wl} is the number of at-risk children who were on the waiting list for Reading Recovery but were not served.

The mean score of each subgroup was weighted by the subgroup's population proportion (over all the districts). All these numbers were summed to arrive at estimates for the statewide means.

$$\bar{X}_{est} = \Sigma[\bar{X}_i (P_i / P_{TOT})] \quad , \qquad \text{(Equation 2)}$$

where $\bar{X}_{est}$ is the estimated statewide mean for X, $\bar{X}_i$ is the mean of X for the i^{th} subgroup, P_i is the population size for the i^{th} subgroup, and P_{TOT} is the total population across all districts among schools with Reading Recovery. (Note that P_{tot} in Equation 1 is the total population *within a district*, P_{TOT} in Equation 2 is the total population *across all districts*.)

Similarly, the variance of each subgroup was multiplied by the subgroup's population proportion to get the sum of squares for that subgroup. These were summed to arrive at estimates for variance for the state as a whole.

$$\sigma^2_{est} = \Sigma[\sigma^2_i (P_i / P_{TOT})] \quad , \qquad \text{(Equation 3)}$$

where σ^2_{est} is the estimated statewide variance, σ^2_i is the variance for subgroup i, P_i is the population size for the i^{th} subgroup, and P_{TOT} is the total population across all districts among schools with Reading Recovery.

An assumption of this technique is that the variance of each subgroup is a reasonable estimate of the variance for the segment of the population that subgroup represents. Although some of the samples representing subgroups were small (sample sizes for the random sample ranged from 4 to 62), there are over 100 districts, and the errors are expected to be random.

Results

The above procedures were used to estimate means and variances for each of the six tests of the Observation Survey administered in the spring of the first grade year. The resulting estimates for the population means and variances on six tests of the Observation Survey are given in Table 1.

Table 1
Means and Variances for Observation Survey Scores

Test	Possible Range	Observed Range	Estimated Statewide Mean	Estimated Statewide Variance	Estimated Standard Deviation
Letter Identification	0 - 54	2 - 54	53.1	3.6	1.9
Concepts About Print	0 - 24	5 - 24	20.2	5.3	2.3
Writing Dictation	0 - 37	2 - 37	34.6	13.7	3.7
Word Test	0 - 20	0 - 20	17.6	11.6	3.4
Writing Vocabulary	0 - 100+[1]	1 - 162	45.9	196.0	14.0
Text Reading	0 - 30	0 - 30	18.8	50.4	7.1

Using the means and standard deviations reported in Table 1, stanines were computed. One quarter of a standard deviation was added to and subtracted from the mean to form the boundaries for stanine 5. Then, one half a standard deviation was added to and subtracted from those

[1]Although there is no established maximum for children's scores on the writing vocabulary task, scores are constrained by the ten minute time limit.

boundaries to form the boundaries for stanines 6 and 4, 7 and 3, and 8 and 2. Stanines 9 and 1 cover all scores higher than stanine 8 and lower than stanine 2, respectively. This procedure is explained in more detail in many measurement texts (e.g., Linn and Gronlund, 1995). The resulting stanine scores are given in Table 2.

Table 2
Resulting Stanine Scores

Observation Survey test	Stanine 1	2	3	4	5	6	7	8	9
Letter ID	0 - 49	50	51	52	53	54	-	-	-
Concepts About Print	0 - 16	17	18	19	20	21	22, 23	24	-
Writing Dictation	0 - 28	29	30, 31	32, 33	34, 35	36, 37	-	-	-
Word Test	0 - 11	12, 13	14, 15	16	17, 18	19, 20	-	-	-
Writing Vocabulary	0 - 21	22 - 28	29 - 35	36 - 42	43 - 49	50 - 56	57 - 63	64 - 70	71+
Text Reading	1 - 6	7 - 9	10, 12	14, 16	18, 20	22, 24	26	28, 30	-

Unlike the Observation Survey, most standardized tests typically cover a wide range of skills, so that, although few students achieve stanine 9 or 1, such stanine scores are possible. Many of the Observation Survey assessments have ceiling effects in the spring. Table 2, which shows the estimated stanine scores for the Observation Survey at the end of the first grade year, indicates this. Many tests do not allow children to demonstrate all the literacy skills they may have acquired.

When scores on an assessment are randomly distributed, the percentage of scores within each stanine follows a pattern. Specifically, 4% of children fall into stanines 1 and 9, 7% fall into stanines 2 and 8, 12% fall into stanines 3 and 7, 17% into stanines 4 and 6, and 20% into stanine 5. Since a simple random sample of Maine first graders does not exist (if it did, there would have been no reason to develop the methods described in this study), these allocations cannot be checked to verify that the stanine scores are appropriately allocated to the raw scores. However, since some schools test all children with the Observation Survey, there is a subpopulation that can serve as a check. The percentages of children within each stanine from schools in which all children were tested are presented in Table 3. It should be noted, however, that most of these children attend small, often rural, schools in Maine, and they therefore do not necessarily represent the state as a whole.

Table 3
Percentages of Children (From Schools that Tested All Children) Within Each Stanine

Stanine	1	2	3	4	5	6	7	8	9
Letter Identification	3%	2%	5%	12%	26%	53%	-	-	-
Concepts About Print	7%	7%	9%	12%	13%	15%	30%	6%	-
Writing Dictation	8%	2%	5%	11%	26%	49%	-	-	-
Word Test	10%	3%	8%	6%	21%	52%	-	-	-
Writing Vocabulary	5%	8%	15%	20%	20%	14%	9%	4%	4%
Text Reading	9%	9%	10%	17%	20%	17%	4%	14%	-
Normal Distribution	4%	7%	12%	17%	20%	17%	12%	7%	4%

Conclusions and Discussion

The Observation Survey has characteristics of both a criterion-referenced and norm-referenced test. On the one hand, it clearly aims to measure skill levels, and it attempts to measure all the major areas of early literacy skill. In this respect it is a criterion-referenced assessment. On the other hand, it is used as a vehicle for assessing the degree to which children are *behind their classmates*, and are therefore at risk for literacy failure. Children scoring in the bottom 20% of their classrooms at the beginning of first grade, as measured by the Observation Survey, are generally recommended for the Reading Recovery program. The aim of Reading Recovery is not simply to help the at-risk child achieve a certain level of skills, but rather to help him or her *reach the average of the classroom*. The Observation Survey is used to assess this progress. In this respect, the Observation Survey is used as a norm-referenced test. It was not designed with all the characteristics of a norm-referenced test, however. For example, it is not designed to capture the full range of literacy abilities first graders could possess. Rather, it focuses on early literacy skills, skills that may not be evident on other assessments or in the classroom.

There are some characteristics of the Observation Survey tests that make stanine computation and interpretation difficult. Many of the tests have floors and ceilings for most children exiting first grade. (They also have floors for many children in Kindergarten and entering first grade.) One of the characteristics of stanines computed on measures with these characteristics is that not all stanine scores are possible. For example, it is impossible for a child exiting first grade to obtain a stanine score above 6 for the Letter Identification task, the Writing Dictation task, or the Word Test. This is logical, though, when the nature of the task is considered. Letter identification is a skill that many

children master in Kindergarten. So many children know the alphabet perfectly by the end of first grade, that it is not a skill on which a child that age can demonstrate extreme above-averageness.

For substantive reasons associated with the domains of six the tests, early literacy practitioners primarily use the Writing Vocabulary and Text Reading assessments for making teaching and programmatic decisions about exiting first graders. The other four assessments are useful only for children who are at an earlier stage of literacy acquisition. The information in Table 3 validates this practice. It is clear that three of the six tests (Letter Identification, Writing Dictation, and the Word Test) would not be able to discriminate well between the skill levels of most children in the spring of the first grade year.

In general, stanines are designed to be used on normally distributed data, and many people may assume a normal distribution when they interpret stanine scores. It is important that the practitioners who use these stanines understand that the data are not normally distributed. Most do, as most practitioners who use the stanines are familiar with the characteristics of the Observation Survey. Still, when the stanines are provided for practitioners in Maine, they are accompanied by a detailed explanation of how they should be interpreted.

Another feature of the Observation Survey that makes stanine interpretation complicated is that some of the assessments are not equal interval scales. This is especially true of the Text Reading measure. Text reading levels are in single-unit gradations from 1 to 10, but after ten, the odd numbers are skipped, although the texts do not appear to get twice as hard with each successive level. The difference between text levels 10 and 12, for example, is not necessarily as large as the difference between levels 2 and 4.

Finally, all the data are from schools in Maine that have implemented the Reading Recovery program. The results given in this paper, therefore, need to be interpreted with this in mind. Many of the schools whose data were included in this study (those with Reading Recovery) may share a curricular and staff development emphasis on early literacy. Students in these schools may have acquired more literacy skills than students in schools that have not targeted early literacy. As a result, the stanines reported here may be higher than those that would have been computed from the scores of all children in the state. The stanines in this paper cannot be taken strictly as stanines for the entire state of Maine, but rather as stanines for those schools in Maine that have implemented the Reading Recovery program.

The methodology in this paper described a technique for estimating population information from a sample not simply randomly drawn, but for which the sizes of the population subgroups were known. This technique was then applied to the computation of stanine scores from a data set associated with the Reading Recovery program. This methodology should be generalizable to other, similar data sets. It is hoped that these procedures will allow researchers to estimate descriptive statistics such as those presented here both for research purposes and for use by practitioners.

References

Clay, M. M. (1993). *An Observation Survey of Early Literacy Achievement*. Portsmouth, NH: Heinemann.

Clinton, W. J. (1997). *State of the Union Address*. Washington, DC.

Linn, R. L. & Gronlund, N. E. (1995). *Measurement and Assessment in Teaching, Seventh Edition*. Englewood Cliffs, NJ: Prentice Hall.

Reading Recovery Council of North America. (1993). *The Executive Summary*. (Technical Report). Columbus, OH: The Ohio State University.

Scott Foresman Reading Recovery Testing Packet. (1979). Glenview, IL: Scott, Foresman and Company.